



## Empirical scoring functions.

# II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model

Christopher W. Murray\*, Timothy R. Auton & Matthew D. Eldridge\*\*

*Proteus Molecular Design Ltd., Beechfield House, Lyme Green Business Park, Macclesfield, Cheshire SK11 0JL, U.K.*

Received 15 November 1997; Accepted 17 February 1998

**Key words:** Bayesian regression, binding affinity prediction, de novo molecular design, protein-ligand complexes, QSAR

### Summary

This paper tests the performance of a simple empirical scoring function on a set of candidate designs produced by a de novo design package. The scoring function calculates approximate ligand-receptor binding affinities given a putative binding geometry. To our knowledge this is the first substantial test of an empirical scoring function of this type on a set of molecular designs which were then subsequently synthesised and assayed. The performance illustrates that the methods used to construct the scoring function and the reliance on plausible, yet potentially false, binding modes can lead to significant over-prediction of binding affinity in bad cases. This is anticipated on theoretical grounds and provides caveats on the reliance which can be placed when using the scoring function as a screen in the choice of molecular designs. To improve the predictability of the scoring function and to understand experimental results, it is important to perform subsequent Quantitative Structure-Activity Relationship (QSAR) studies. In this paper, Bayesian regression is performed to improve the predictability of the scoring function in the light of the assay results. Bayesian regression provides a rigorous mathematical framework for the incorporation of prior information, in this case information from the original training set, into a regression on the assay results of the candidate molecular designs. The results indicate that Bayesian regression is a useful and practical technique when relevant prior knowledge is available and that the constraints embodied in the prior information can be used to improve the robustness and accuracy of regression models. We believe this to be the first application of Bayesian regression to QSAR analysis in chemistry.

### Introduction

In a previous paper [1], we developed a fast empirical scoring function from a training set of 82 ligand-receptor complexes for which the binding geometries and binding affinities were known. The approach adopted was similar to the work of others [2–4] but employed a larger training set and different func-

tional forms. The scoring function contained only 4 terms (not including the intercept), and each term and coefficient had a direct physical interpretation, i.e., hydrogen bonding, metal binding, lipophilic contact and flexibility penalty. Extensive testing of the function on different choices of training set and application to test sets indicated that the coefficients are reasonably stable to the introduction of new complexes into the training set. The scoring function had a cross-validated error of 8.68 kJ mol<sup>-1</sup> (approximately 1.5 orders of magnitude in the binding affinity).

\*To whom correspondence should be addressed.

\*\* Present address: E.M.B.L. Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

The previous paper pointed out a number of weaknesses with the overall approach. Perhaps the most serious is that the training and test sets only contain information on molecules which have measurable binding affinities. Real applications of the scoring function will be for candidate molecular designs which are not guaranteed to bind. The candidate designs may contain features which inhibit binding and are poorly represented in the training set of molecules with measurable binding affinities. Additionally, the binding mode proposed for molecular designs may be in serious error. Clearly if the proposed binding mode is wrong, any empirical estimate of binding affinity which is based on that binding mode will also be in error. The situation is made worse because the typical selection of designs chosen for synthesis is expected to be enriched with molecules with plausible binding modes and high predicted scores. It is therefore to be anticipated that the selection process will identify false positives (i.e., molecules with significantly poorer binding affinities than predicted) reflecting the fact that some facets which are unfavourable for binding are not represented adequately in the training set.

This paper examines a set of molecular designs produced by our synthetically constrained *de novo* design program, PRO\_SELECT [5, 6]. These molecules were designed to bind to thrombin and one of the criteria for choosing synthesis candidates was their predicted empirical score. The list of candidate designs and their associated binding modes therefore provide a good and a different test of the applicability of the scoring function. To our knowledge this is the first extensive test of an empirical scoring function of this type with a set of molecular designs which were subsequently synthesized.

As indicated above, it is anticipated that molecular designs chosen on the basis of an empirical scoring function will contain false positives when the function has been primarily trained on crystallographically determined ligand-receptor complexes. This of course does not negate the usefulness of the scoring function since the requirement of a good score will still considerably reduce the molecule space needing to be sampled when synthesis candidates are chosen. However it does mean that during the normal operation of a structure based drug design project which employs empirical scores, it will be necessary to amend and update the scoring function in the light of assay data. In this way a progressively more reliable scoring function can be derived as the drug discovery project proceeds. This paper illustrates the use of follow-up

QSAR in providing a more reliable function and in interpreting the discrepancies between the predicted and experimental binding affinities.

Follow-up QSAR would involve using the empirical scoring function or its components as descriptors, together with any new descriptors which a scientist thought important in analysing the data. However, a normal application of QSAR would be restricted to one set of incoming data; it is difficult to simultaneously take account of the prior information contained in the previous training set and its associated regression. What is required is a method which uses the prior information in the new regression – a method which matches the new data whilst monitoring the performance of the new regression on the old training set. This reflects the way chemists and modelers feel about new data on drug discovery projects, i.e., there is a collection of existing data on related enzymes or molecules, and new data is interpreted and examined in terms of that prior information. A method that provides the proper mathematical framework for using prior information to model new information is Bayesian statistics [7]. This paper uses Bayesian regression on an empirical scoring function with additional parameters to model data on a set of synthesised and assayed molecular designs. The prior information is derived from the training set of ligand-receptor complexes used in the original derivation of the scoring function [1]. Bayesian regression is also used in an internal test on serine proteases present in the ligand-receptor complex database. To our knowledge this is the first time that Bayesian regression has been used for chemical applications of QSAR.

The next section describes the methods and materials used in this paper. In particular it describes the training set of molecular designs and how the geometries for those designs were prepared. It gives a quick overview of the parameters used in the regression although most of these are described in the previous paper. It also describes the Bayesian approach to regression which is adopted in this paper. The section that follows this gives the results and this is followed by a discussion of the results. The final section gives the conclusions.

## Methods

### Training set

The molecules used in this study were produced by an application of PRO\_SELECT to thrombin inhibitors [6]. The designs were based on PPACK, a known proline-containing inhibitor of thrombin. All designs took the form given in Figure 1, with carboxylic acids or sulphonic acids being chosen at the A position and amines being chosen at the B position. The proline template occupies the central P pocket of thrombin; the A component contains a hydrophobe which occupies the D pocket; and the B component contains functionality to enter the S1 subsite and to hydrogen bond at the bottom of that pocket, usually with an aspartate group (Asp-189). One of the criteria for choosing synthesis candidates was their predicted empirical score using an implementation of Böhm's function [2]. Although there are differences between our current scoring function and the one used to help choose synthesis candidates, we believe that false positives in our implementation of the Böhm score are very likely to coincide with false positives in the current scoring function. The binding mode of the inhibitor originally produced by PRO\_SELECT was based on one of two positions generated for the receptor and proline moiety: the first was derived directly from the crystallographic structure for PPACK taken from the Brookhaven PDB entry 1DWE [8]; and the second from some modelling of a non-covalently bound analogue of PPACK based on the PDB entry, 1DWC [8]. This latter model was employed in the modelling of the designs used in this study. A model for trypsin was derived from the Brookhaven entry, 1PPC [9] which is in the same reference frame as 1DWC. A good initial binding mode for each design in trypsin is generally obtained by using the coordinates from the thrombin binding mode.

The initial binding mode for each molecule in each enzyme was treated by the following procedure using Discover [10] with the CFF95 Forcefield [11]. Local refinement of enzyme:ligand complexes was accomplished by conjugate gradient minimization (to a maximum gradient of  $0.1 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ ), treated *in vacuo* with a distance dependent dielectric, and a non-bonded cut-off of  $20 \text{ \AA}$ . In order to sample local minima, a short molecular dynamics simulation was performed (10 ps at 150 K, saving snapshots every 2 ps, with subsequent minimization of each snapshot). In this study, only the first snapshot was used dur-

ing the regression but the results are similar when, for example, the fifth is used. The molecular dynamics protocol only allows local rearrangement of the enzyme:ligand interface and the binding mode chosen will be similar to the one produced directly by PRO\_SELECT. There is no guarantee that this binding mode is correct although it is likely to be a good approximation for the better inhibitors.

There are 31 molecules considered in this study and the binding affinities of the molecules against trypsin and thrombin are reproduced in Table 1 [6]. Figure 1 gives the structure of the components of the molecules from Table 1. No other molecules were synthesised in this library except one molecule which was omitted for technical reasons (it is not an outlier) and one molecule which did not have a measurable affinity for trypsin (it contained a neutral S1 donor group, so activity is expected to be low from the analysis in this paper). The variation in activities is low, especially in the case of thrombin. Traditional QSAR would find it very difficult to arrive at a good model using this training set. However, the training set is a fair reflection of the type of molecules produced in the early phases of structure based drug discovery projects.

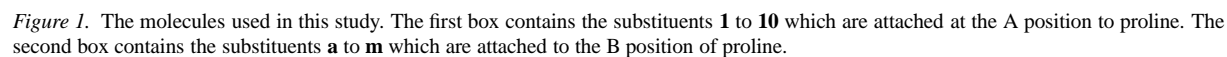
In addition to the training set of molecular designs, we also need relevant prior information. This comes from a set of 82 crystallographically determined protein-ligand complexes for which a binding affinity is known. It contains 17 aspartic protease, 15 serine protease, 15 metalloprotease, 16 sugar-binding protein and 19 other complexes. This paper will also examine the use of Bayesian regression to fit the 15 serine proteases using the other 67 complexes as prior information. The database and its construction are described in detail in another paper [1].

### The regression terms

The empirical scoring function used in this work [1] can be written in the form:

$$\begin{aligned} \Delta G_{\text{binding}} = & \Delta G_0 + \Delta G_{\text{Hbond}} \sum_i g_1(\Delta r) g_2(\Delta \alpha) \\ & + \Delta G_{\text{metal}} \sum_a m f(r_{aM}) \\ & + \Delta G_{\text{Iipo}} \sum_{iL} f(r_{iL}) \\ & + \Delta G_{\text{rot}} H_{\text{rot}} \end{aligned} \quad (1)$$

The  $\Delta G$  coefficients are:  $-5.48$ ,  $-3.34$ ,  $-6.03$ ,  $-0.117$  and  $2.56$ , respectively. The first term is simply the intercept and it is the least statistically significant term in the regression. The second term scores the number and quality of the hydrogen bonds in a similar



**Figure 1.** The molecules used in this study. The first box contains the substituents **1** to **10** which are attached at the A position to proline. The second box contains the substituents **a** to **m** which are attached to the B position of proline.

Table 1. Binding affinities of compounds in the test set, presented as  $pK_i$  (i.e.,  $-\log K_i$ ). The molecules are of the form A-Pro-B, and the structures are given in Figure 1

Compound	$pK_i$ thrombin	$pK_i$ trypsin
<b>1a</b>	6.95	6.61
<b>1b</b>	6.08	6.71
<b>1c</b>	5.56	5.06
<b>1d</b>	4.78	2.79
<b>2a</b>	6.29	6.02
<b>2b<sup>a</sup></b>	5.22	6.66
<b>2c<sup>a</sup></b>	4.96	5.38
<b>2d</b>	4.57	2.97
<b>3a</b>	7.44	6.22
<b>3b</b>	6.53	6.02
<b>3c</b>	5.88	4.85
<b>3d</b>	5.24	3.11
<b>4a</b>	5.84	5.88
<b>4b</b>	5.31	6.10
<b>4c<sup>a</sup></b>	4.91	5.28
<b>4d</b>	4.09	2.49
<b>1e<sup>a</sup></b>	5.27	4.80
<b>1f</b>	4.10	1.94
<b>1g</b>	6.15	4.23
<b>1h</b>	5.14	5.42
<b>1i</b>	6.16	5.83
<b>1j</b>	5.40	3.23
<b>1k</b>	3.00	2.26
<b>1l</b>	2.86	2.88
<b>1m</b>	2.76	1.89
<b>5a</b>	5.54	7.03
<b>6a</b>	6.56	6.13
<b>7a</b>	5.80	6.17
<b>8a</b>	6.27	6.38
<b>9a<sup>a</sup></b>	4.11	5.45
<b>10a<sup>a</sup></b>	4.20	4.44

<sup>a</sup> 0.4 Units have been added to the activities for the enzymes since these molecules were tested as crude samples.

way to Böhm [2], but no distinction is made between charged and uncharged interactions, and water mediated hydrogen bonds are also scored. (The inclusion of water-mediated hydrogen bonds is justified for the crystallographically determined protein-ligand complexes in the original training set but is not relevant to the modelled thrombin designs which have been modelled against dry enzyme structures.) The third term scores contacts between metals and appropriate hetero-atoms in the ligand. In this application to serine

proteases, the metal term will never contribute to the binding affinity. The fourth term is an estimate of the interaction between lipophilic atoms in the ligand and lipophilic atoms in the receptor. The term has a simple contact-based form and is reasonably long range so that atoms within about 7 Å contribute to the score. The final term is an estimate of the entropic penalty associated with flexible ligands. The term considers rotatable bonds whose rotations are frozen as a result of ligand binding, and scores the bonds according to the lipophilicity of the atoms on either side of the bond. The exact form of the terms and their physical interpretation are given elsewhere [1].

An extra term is used in Bayesian regression on trypsin. It counts the charged hydrogen bond contacts to the critical aspartate group (Asp-189) at the bottom of the S1 pocket and stores it as a separate column for the regression. Aliphatic amines and guanidines are judged capable of giving charged contacts, so that charged hydrogen bond contributions occur only for components **a**, **b**, **c**, **e**, **h** and **i** given in Figure 1. The contact term is of the same form as the one used in the full hydrogen bond term [1]. The charged hydrogen bond term is adjusted so that it has zero mean across the 31 molecules studied. Study of the ligand-receptor complexes in the training set for which Equation (1) was established, gives us prior knowledge that separating charged ionic terms from the hydrogen bond term does not help regressions on the training set, but analysis of the results for the trypsin test set indicates that for interactions with Asp-189 in trypsin, *but not thrombin*, a charged hydrogen bond is important.

An extra term is also added in the Bayesian regression on thrombin. Here it is obvious that the charged ionic term would not help the agreement with experiment because several charged B components have poor activity and some uncharged moieties have good activity. There does however seem to be a preference for the guanidine-containing B component (agmatine, labelled '**a**' in Figure 1). We introduce an indicator variable for agmatine as a column in the Bayesian regression, and again arrange for the column to have zero mean. It is with reluctance that we introduce such a specific indicator variable since we would prefer to employ a more general and physically based term. However, investigation of various physical terms yields regressions which appear valid but are essentially employing the physical term as a method of separating out agmatine-containing molecules from the others. This is because with many physical terms, agmatine is an outlier. In such a case, we believe it

is safer to use a true indicator variable, and avoid the danger of relying on potentially spurious physical explanations. Additionally it proved to be necessary to remove some outliers from the regression in order to get good results. The molecules omitted were **9a**, **10a**, **1k**, **1l** and **1m**, the reasons for their omission are discussed later.

### Bayesian regression

Classical linear regression gives us two options to making QSAR predictions for activities of ligands against a specific enzyme.

(1) Use an equation based on general purpose descriptors, and fitted to a database of disparate ligands and enzymes.

(2) Develop an equation specifically for this enzyme, using measured activities of ligands against this enzyme only.

The weakness of 1 is that it does not make any effective use of available data for the particular target enzyme. This will probably result in model predictions which are useful as a crude pre-screen, but not accurate enough to reliably rank activity of similar structures. The weakness of 2 is that the available database of compounds assayed against the target enzyme will usually be too small to adequately cover the chemical space of potential ligands. An enzyme specific QSAR equation is unlikely to provide useful predictions of activity of compounds which are not similar to those already tested. However, such an equation may provide predictions of high accuracy for conservative changes of structure away from training set of compounds. There is more flexibility in choosing descriptors for a type 2 equation. For example, indicator variables can be used if it is seen that particular sub-structures are correlated with activity, even if the mechanism by which this sub-structure interacts is uncertain.

Our motivation in exploring the use of Bayesian regression is that it should provide us with a mathematically sound methodology in which to develop hybrid QSAR models, which, we hope, may retain the robustness of general purpose equations while allowing enzyme specific data to be used to improve model accuracy. The methodology we have used allows us, effectively to augment the general purpose training set with enzyme specific data, and to introduce new descriptors which are specific for the target enzyme.

Bayesian regression incorporates these components:

**Prior** which gives prior knowledge of the distribution of model parameters (coefficients) and error terms.

**Data** a set of quantities to be fitted (e.g., enzyme specific activity measures) and descriptor values.

**Likelihood** A statistical model for the likelihood of observing these activities, given estimates of the model parameters.

These data are then used with Bayes equation to provide a posterior distribution of model parameters, which incorporates all of our information about the model parameters. Summaries of this distribution such as marginal expectations and variances can then be calculated and used to assess the accuracy and validity of the model, and to make predictions of activity for new structures.

### Implementation

Our implementation of Bayesian regression tools closely follows the theory of linear models described in [7] (Chapter 9).

We use independent prior distributions for model coefficients ( $\beta$ ) and the error variance ( $\sigma^2$ ), as described in [7] (paragraph 9.62), as it is considered closer to our subjective prior and simplifies specification of the parameters.

The prior distribution of the regression coefficients (including the intercept term)  $\beta$  is assumed to be a multivariate normal distribution with mean  $\beta_{\text{prior}}$  and variance  $\mathbf{I}_{\text{prior}}^{-1}$ , where  $\mathbf{I}_{\text{prior}}$  is the prior information matrix.

$$f(\beta) \propto \exp\left(-\frac{1}{2}(\beta - \beta_{\text{prior}})' \mathbf{I}_{\text{prior}}(\beta - \beta_{\text{prior}})\right) \quad (2)$$

The prior distribution of  $\sigma^2$  is assumed to be an inverse chi-squared distribution, and independent of the prior distribution of the coefficients  $\beta$ .

$$f(\sigma^2) = \frac{e^{-\frac{1}{2}z} z^{\frac{1}{2}v-1}}{2^{\frac{1}{2}v} \Gamma(\frac{1}{2}v)} \quad (3)$$

where

$$z = \frac{\sigma_{\text{prior}}^2}{\sigma^2} \quad (4)$$

and  $v$  and  $\sigma_{\text{prior}}^2$  are chosen so that the 5 and 95 percentile points of the distribution are equal to the end-points of a user-specified 'plausible range'.

The likelihood function is the same as for classical linear regression

$$f(\mathbf{y}|\beta, \sigma^2) = (\pi\sigma^2)^{-n/2} \exp\left(-(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)/(2\sigma^2)\right) \quad (5)$$

The posterior distribution of  $\beta$  and  $\sigma^2$  is then given by Bayes' theorem

$$f(\beta, \sigma^2 | \mathbf{y}) \propto f(\mathbf{y} | \beta, \sigma^2) f(\beta) f(\sigma^2). \quad (6)$$

We have found it informative to summarize this distribution using the mean and variance of the marginal distributions of  $\beta$  and  $\sigma^2$ . These can be calculated rapidly using one-dimensional quadratures [7]. Quadratures were performed using routine D01BCF from the NAG library [12].

#### *Using Bayesian regression*

The theory behind linear Bayesian regression is standard, however it may not be immediately obvious to the general reader how Bayesian regression is used in practice and in particular, how the user initiates and runs the regression.

In our implementation, the input to the Bayesian regression is made via an Xwindow interface and requires:

- the data to be fit. For example, the activity data against trypsin for the molecules in the test set.
- The descriptors used to fit the data. In our case, this might be the values of each of the terms in equation 1 for all the molecules in the test set.
- The prior estimate of the coefficients ( $\beta$ ). In our case, these might be the coefficients from a previous regression on a more general and extensive training set.
- The information matrix. This is derived from an estimate of the relative errors in the coefficients. A suitable information matrix might be the inverse of the covariance matrix from a related regression. The relative sizes of the diagonal of the information matrix determine the relative freedom to vary the coefficients in the regression (larger relative values imply more highly constrained coefficients).
- An estimate of the plausible range of rms error anticipated for the new data. If the estimated errors are small then there will be greater freedom to vary the coefficients, since this implies the new data should be fit more accurately.

Since it is usual to derive the prior information from a previous regression, it is convenient to run that regression from the Xwindow interface while specifying null prior coefficients and a null information matrix. This amounts to running a normal multiple linear regression. The inverse of the covariance matrix will be produced and can be saved together with the regression coefficients. These can then be input directly or in a manipulated form as the prior information into

a Bayesian regression with new data. The Bayesian regression is therefore easy to use and the user can experiment with the various parameters in an attempt to produce an improved model of the data.

#### *The regressions and their assessment*

Bayesian regression can be thought of as a way of performing a range of QSAR analyses between two extremes. There exists prior information in the form of a previous regression on relevant set of prior data – the training set of complexes used to fit the original scoring function. There is also a set of data which is to be fit, which here might be the trypsin or thrombin data for the molecular designs. One extreme of analysis, is to assume the prior information is correct and that the new data can not improve on this model. This amounts to using the prior regression coefficients predictively on the new data and looking at the results. This extreme will be referred to here as *Model(0)* since no weight is given to the data in the analysis. The other extreme of analysis will be referred to as *Model( $\infty$ )* since in this extreme, the data is given infinite weight in comparison to the prior information. This amounts to performing a multiple linear regression on the data without considering the prior at all. In between these two extremes are an infinite number of Bayesian regressions in which differing weights can be given to the data. We will consider two alternative models in addition to *Model(0)* and *Model( $\infty$ )*. *Model(1)* will use an information matrix derived from the inverse of the covariance matrix from the related prior regression together with its estimates for the coefficients. *Model(2)* is similar to *Model(1)* but uses an information matrix with more freedom to vary the coefficients; the diagonal of the covariance matrix is doubled prior to inversion. For both *Model(1)* and *Model(2)*, the plausible range of rms error chosen is 3.5–10.0 kJ mol<sup>-1</sup>, since this was thought reasonable for our data.

The methods used for assessing the regressions in this work are fairly standard. They fall into three categories. The first concerns how well the data is fit and the quantities looked at are the rms residuals of the fitted data, the average residuals, the  $R^2$  value of the fitted data and the correlation coefficient,  $r_P$ . The second category is concerned with internal robustness in the data and looks at the leave-one-out cross validated quantities,  $R_{cv}^2$ ,  $SPRESS$  and the average cross validated residual. The third category considers how well prior information is fit which is another (perhaps more se-

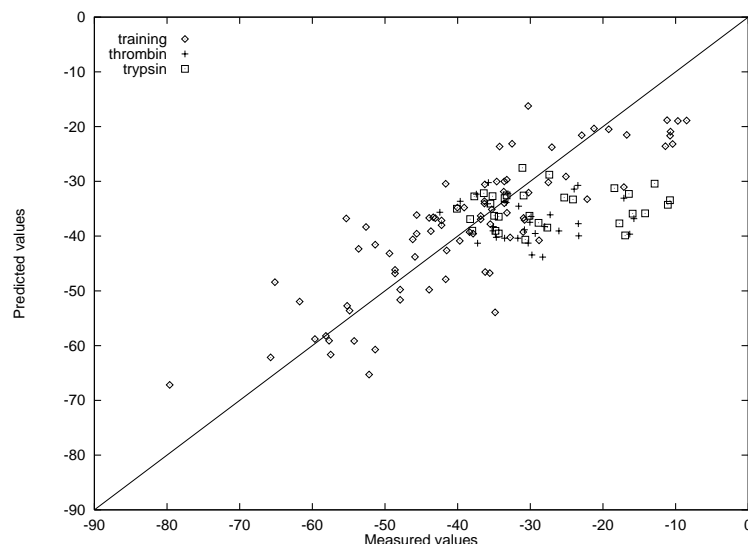


Figure 2. A scatter plot showing the fitted binding affinities versus the experimental binding affinities in  $\text{kJ mol}^{-1}$ , for the thrombin and trypsin test sets superimposed on the original training set. The predicted values were obtained using Equation (1).

vere) measure of robustness. The quantities examined here are the rms residuals and average residuals of the 67 members of the protein-ligand database [1] which *are not* serine proteases. The use of only the non-serine protease parts of the prior information provides clearer interpretation for this metric.

Bayesian regressions will be run on 3 different sets of data. The first set is the 15 serine proteases in the protein-ligand database using prior information derived from the other 67 complexes. This is meant as an example to show how the Bayesian regression works. The second regression is based on the trypsin data. The third regression is based on the thrombin data although in this case the data set is of such low quality that it is difficult to derive a predictive model. Nevertheless the thrombin data set is typical of the type of data coming from initial synthesis rounds in a structure based drug design project and therefore is a relevant application of the Bayesian approach.

## Results

### *Predictions on the molecular designs*

First the performance of Equation (1) in predicting the trypsin and thrombin data is presented. The rms residual for the trypsin results is  $11.6 \text{ kJ mol}^{-1}$  and the mean residual is  $+7.2 \text{ kJ mol}^{-1}$ . The rms residual for the thrombin results is  $10.4 \text{ kJ mol}^{-1}$  and the mean residual is  $+7.1 \text{ kJ mol}^{-1}$ . Both predictive sets

therefore show a significant tendency to be overpredicted. As mentioned earlier, overprediction is likely to occur since the training set and the associated regression are not well suited to predicting the binding affinity of molecules which contain unfavourable aspects of binding. However, despite this deficiency, the rms residuals compare quite favourably with the cross-validated error for the original training set ( $8.7 \text{ kJ mol}^{-1}$ ). The relative performance of the training set and the test sets are illustrated in Figure 2.

Figures 3 and 4 give individual scatter plots of the molecular design set against trypsin and thrombin respectively. It is clear from these plots that the correlation between the predicted and actual binding affinities is very poor (i.e., does not exist). However it must be remembered that the test sets are especially difficult. The molecules were all chosen to have good predicted binding affinities against thrombin and also to possess as much diversity as possible within the confines of the design goals [6]. As a result, the spread in predicted activity is small, especially for thrombin, and given the known inaccuracies in prediction (i.e.,  $8.7 \text{ kJ mol}^{-1}$  as the cross-validated error) poor correlations must be expected. It is apparent that the method is not good at predicting the ordering of activity amongst a collection of similar analogues with plausible geometries. Again this short-coming is anticipated since similar observations can be made for some sets of analogues in the original training set [1]. This deficiency does not negate the usefulness of the



scoring function as a broad screen of potential designs for synthesis or testing.

to thrombin is also supported by extensive SAR in a recent paper on thrombin inhibitors [14].

The inclusion of an ionic term would not improve the correlation for thrombin, instead it is noticeable that agmatine (**a**) in the B position gives relatively strong binding compared to other charged moieties. One possible reason for this is the high  $pK_a$  of agmatine compared to the other bases. An exception to this observation are compounds **9a** and **10a** which are sulphonamides. Unlike the other A components, these sulphonamides are not predicted to form a hydrogen bond with the carbonyl of Gly-216, and are predicted to occupy the D pocket in a different orientation. If either of these factors is particularly important then this would account for their poor activities compared to the other agmatine containing compounds. The sulphonamide compounds are omitted in subsequent analyses. Figure 4 also contains three other outliers (**1k**, **1l** and **1m**) omitted in the later Bayesian regressions on thrombin. These compounds were synthesised in the light of the unexpectedly good thrombin binding and the thrombin/trypsin selectivity of **1g**. The compounds were specifically chosen to explore greater diversity and probe other potential hydrogen bonds in the S1 pocket. In particular **1k** and **1m** were made even though it was suspected that they might be un-ionised in the assay (at the pH conditions of the experiment). If this were the case, the pyridyl nitrogen will probably form poor electrostatic contacts with oxygens in the S1 pocket leading to poor activity. Such negative

contributions to the binding affinity will be difficult to predict with the empirical scoring function. **1h** is an odd molecule and is considerably larger than the other S1 binders; it is possible that it is too big for the S1 pocket and that the predicted binding geometry is incorrect. The above discussions highlight the need for more QSAR to improve the prediction and interpret the activity of the molecular designs.

### Bayesian regression on serine proteases

reduction in the tendency to overpredict by adjusting the intercept. The intercept is the easiest term to change since in the Model(0) regression, it is the least statistically significant of the coefficients. Superior performance for the serine proteases is at the expense of less good performance on the other complexes. This is precisely what one would expect to see. The Bayesian regressions allow one to tailor the performance against the serine proteases whilst constraining the regression to still give acceptable performance against useful and relevant prior information. This should produce models which are more transferable and robust than Model( $\infty$ ) and more accurate for the data being fit than Model(0).

### Bayesian regression on molecular designs

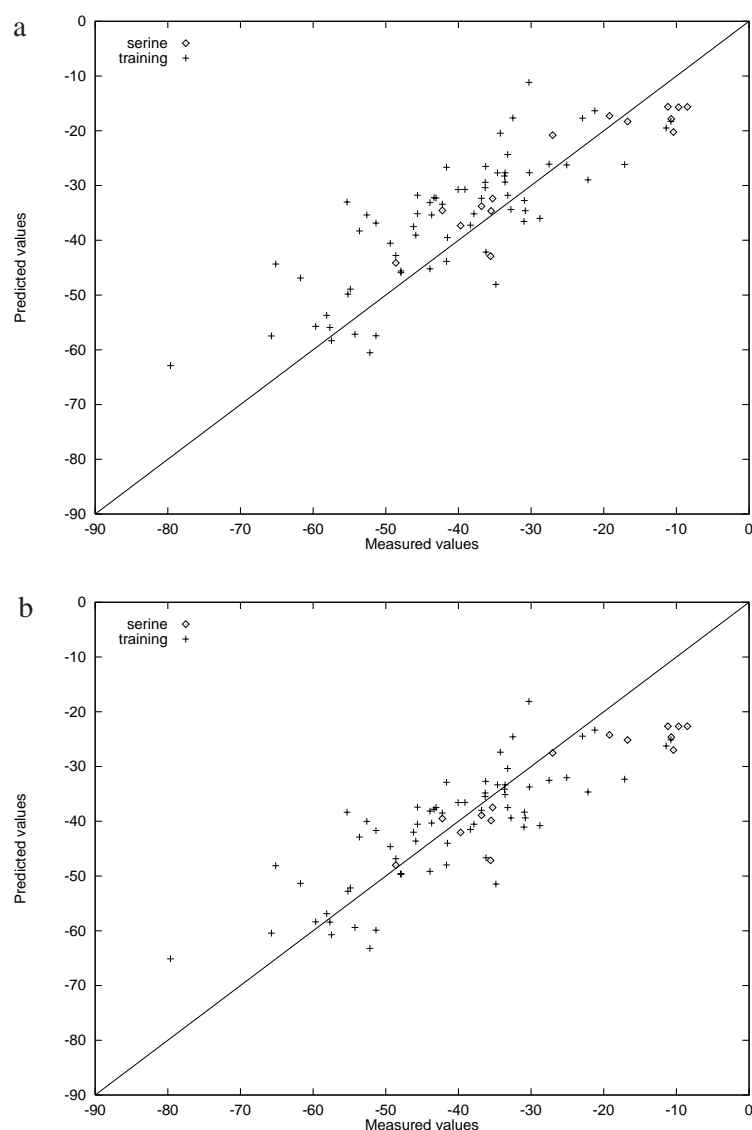
Table 3 shows the results for fitting the affinities of the molecular designs against trypsin. The prior in-

*Table 2.* The regression coefficients and information on fitting using different regression models for the binding affinities of 15 serine protease-ligand complexes. The prior information is derived from 67 non-serine protease protein-ligand complexes. Positive mean residuals indicate systematic over-prediction

		Model(0)	Model(1)	Model(2)	Model( $\infty$ )
Coefficients	Intercept	−10.8	−3.72	−2.23	5.31
	H Bond	−3.05	−3.44	−3.32	−3.95
	Metal	−5.04	−6.36	−5.62	0.00
	Lipophilic	−0.108	−0.121	−0.122	−0.182
	Flexibility	2.65	2.56	2.85	5.00
Fit to data	Rms residual	9.10	6.27	5.51	4.63
	Mean residual	6.82	3.26	0.92	0.00
	Correlation ( $r_P$ )	0.933	0.929	0.933	0.940
	$R^2$	—	0.785	0.834	0.883
Cross-validated	$sPRESS$	—	8.09	7.71	8.64
fit to data	Mean residual	—	3.35	1.00	0.55
	$R^2_{CV}$	—	0.762	0.784	0.729
Fit to 67 protein-ligand complexes	Rms residual	7.74	8.07	9.24	12.33
	Mean residual	0.00	−1.22	−4.88	−6.90

*Table 3.* The regression coefficients and information on fitting using different regression models for the binding affinities against trypsin of 31 molecular designs. The prior information is derived from 82 protein-ligand complexes. Positive mean residuals indicate systematic over-prediction

		Model(0)	Model(1)	Model(2)	Model( $\infty$ )
Coefficients	Intercept	−5.48	−1.59	−4.25	−17.2
	H Bond	−3.37	−3.28	−2.68	−1.17
	Metal	−6.03	−7.85	−6.17	0.0
	Lipophilic	−0.117	−0.114	−0.105	−0.0652
	Flexibility	2.56	2.18	2.66	2.13
	Asp-189 Ionic	0.00	−4.79	−5.24	−5.52
Fit to data	Rms residual	11.62	5.66	4.20	4.01
	Mean residual	7.15	3.60	0.19	0.0
	Correlation ( $r_P$ )	0.128	0.876	0.885	0.895
	$R^2$	—	0.604	0.782	0.801
Cross-validated	$sPRESS$	—	6.65	5.07	5.13
fit to data	Mean residual	—	3.75	0.22	0.11
	$R^2_{CV}$	—	0.559	0.744	0.737
Fit to 67 protein-ligand complexes	Rms residual	7.93	8.82	11.97	15.61
	Mean residual	−0.91	−3.26	−9.00	−9.74



**Figure 5.** Scatter plot showing the fitted binding affinities versus the experimental binding affinities in  $\text{kJ mol}^{-1}$ , for the serine proteases and the other complexes in the training set. (a) The fitted values were obtained from Bayesian Model(2) of Table 2. (b) The fitted values were obtained from Bayesian Model(0) of Table 2. This is the model where the regression obtained from the training set without the serine proteases is used directly as a prediction of the binding affinities for the serine proteases.

formation was derived from the full regression which yielded Equation (1). The descriptors were augmented by a term indicating the strength of ionic interaction of the designs with Asp-189, and no prior information on this term was used (i.e., it is not constrained to any original value in the regression). The Model(0) results are equivalent to predictions using Equation (1) and have been discussed above – there is very little correlation, the rms residuals are very poor and there is a clear tendency to over-predict. Use of Bayesian regression

with the new variable greatly improves the fit. The Model(1) results are plotted in Figure 6 and there is still a tendency to over-predict. Attaching more weight to the trypsin data in Model(2) further improves the fit and the cross-validated results primarily through the reduction in overprediction as illustrated in Figure 7. However the fit to the 67 non-serine proteases is quite badly affected and the rms residual worsens by  $3 \text{ kJ mol}^{-1}$ . This is a good test of robustness and potential transferability, since the prior information

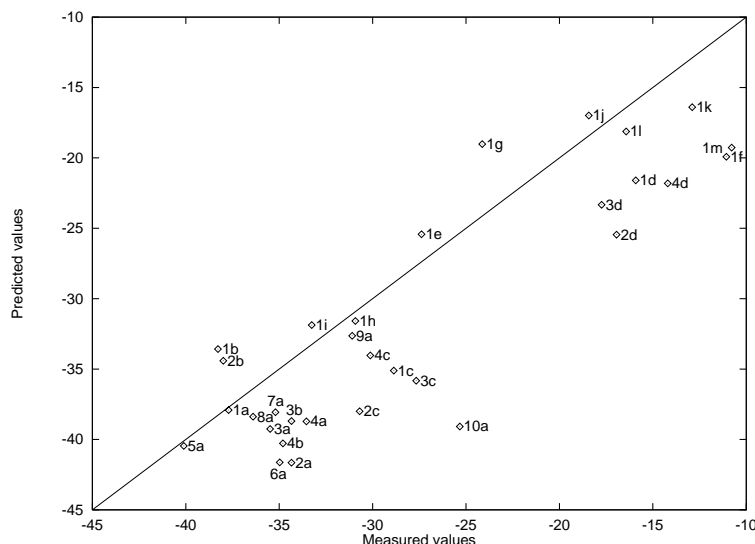


Figure 6. A scatter plot showing the fitted binding affinities for the trypsin data set versus the experimental binding affinities in  $\text{kJ mol}^{-1}$ . The fitted values were obtained using Bayesian Model(1) of Table 3.

should be highly relevant when the scoring function is applied to new molecular designs. The judgement as to whether Model(1) is preferred over Model(2) depends on how well one wishes to fit the trypsin results at the expense of mis-fitting the prior information. In our view both Model(1) and Model(2) are potentially useful models. Model( $\infty$ ), however, is not useful since there is considerable over-fitting. There is not enough diversity in the trypsin data and the descriptors to fit the regression equation well. The large change in the coefficients indicates the lack of robustness, as do the poor predictions on the prior information. Interestingly, the leave one out cross-validated results are respectable, showing the danger of relying too heavily on these quantities rather than employing larger cross-validation sets or prediction on separate test sets.

The Bayesian regression was applied to the thrombin data and for the reasons outlined above **9a**, **10a**, **1k**, **1l** and **1m** were omitted from the analysis. An agmatine indicator variable was used as an additional descriptor. The results are shown in Table 4. Here the poor diversity of the data and the descriptors prohibits the production of models with good correlations. This is particularly clear from the Model( $\infty$ ) results where straightforward regression produces a very poor model. An equivalently poor model can be obtained using just an intercept and an agmatine indicator variable. A good test of the robustness of the model is obtained from the rms residuals against the 67 non-serine protease-ligand complexes. Model(0) gives rea-

sonable rms residuals for the thrombin data but poor correlations – again indicating the lack of variation in the predicted activities. Model(1) is a substantial improvement and yields some correlation and very good rms residuals, as illustrated in Figure 8. The cross-validated  $R^2$  is still poor though – primarily because of a systematic over-prediction. It can be seen that the preference for agmatine as a component is strong in this data set. Obviously this would not be transferable to general compounds outside this series. Nevertheless, the regression still provides a useful analysis of the data. As mentioned above, attempts at using more general descriptors were not helpful since the preference for agmatine is very strong and agmatine is an outlier in the space of many plausible descriptors. More compounds need to be synthesised before more transferable models can be formulated. The tendency to over-predict is removed in Model(2) where more weight is placed on the thrombin data in the regression. As with the trypsin data, this improvement in the fitting is at the expense of poorer fitting of the prior information. Figure 9 gives the scatter plot of predicted versus experimental binding affinity for Model(2).

## Discussion

The first requirement in the application of structure based drug design is a reliable way of selecting molecules to be synthesised using the structural information of the target as a basis. The empirical

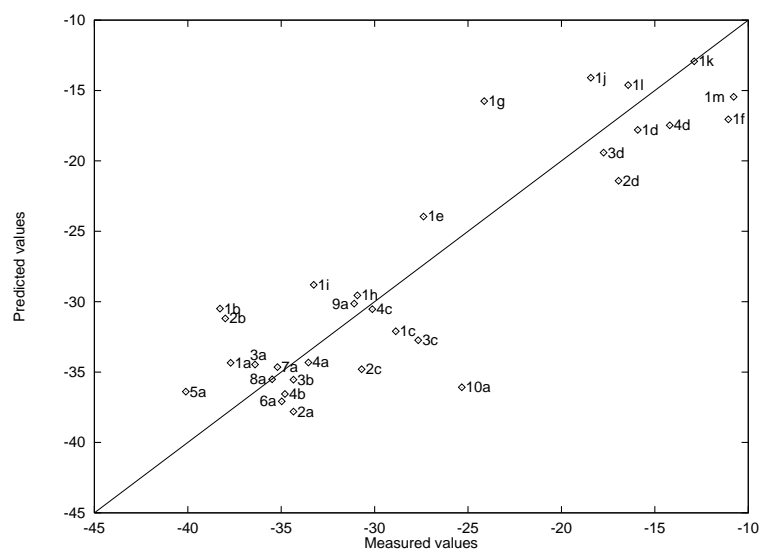


Figure 7. A scatter plot showing the fitted binding affinities for the trypsin data set versus the experimental binding affinities in  $\text{kJ mol}^{-1}$ . The fitted values were obtained using Bayesian Model(2) of Table 3.

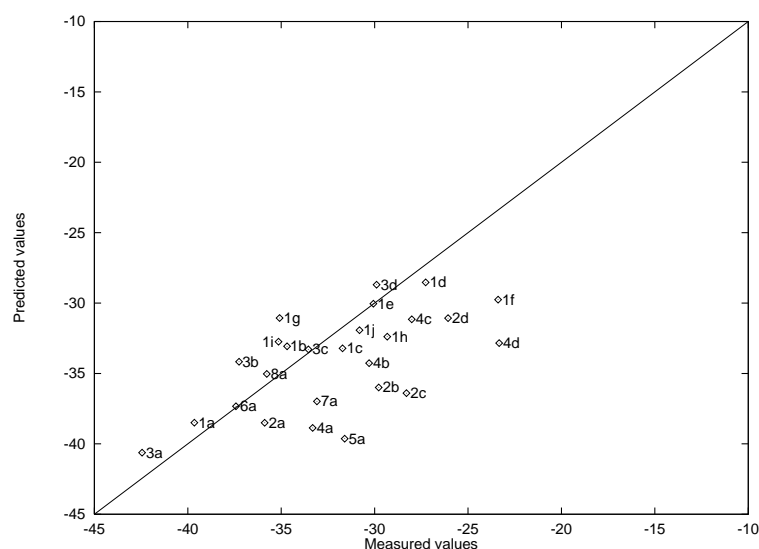


Figure 8. A scatter plot showing the fitted binding affinities for the thrombin data set versus the experimental binding affinities in  $\text{kJ mol}^{-1}$ . The fitted values were obtained using Bayesian Model(1) of Table 4.

approach suggested by Böhm [2] and built upon by others [1, 3, 4] represents a practical approach to assessing a wide diversity of different molecular designs and forms the basis for a semi-quantitative screening of the designs. The methods have been shown to work well on a fairly diverse set of crystallographic complexes for which the binding affinity and binding geometries are known. However this paper represents the first substantial, practical test of the methods on a set of molecular designs – many of which proved

to be poor binders. As anticipated, the empirical scoring function tested in this paper proved to give poor predictions when applied to some of the molecular designs. This can be traced back to the fact that molecules which bind poorly are not well represented in the training set. We believe that other empirical scoring functions of the same type will suffer from similar difficulties. Another source of potential error is that the binding modes used to calculate the empirical scores may be incorrect. The problems are exemplified by

Table 4. The regression coefficients and information on fitting using different regression models for the binding affinities against thrombin of 26 molecular designs. The prior information is derived from 82 protein-ligand complexes. Positive mean residuals indicate systematic over-prediction

		Model(0)	Model(1)	Model(2)	Model( $\infty$ )
Coefficients	Intercept	-5.48	-3.48	-4.48	-58.7
	H Bond	-3.37	-3.60	-3.30	-2.77
	Metal	-6.03	-7.09	-6.02	0.0
	Lipophilic	-0.117	-0.103	-0.103	+0.099
	Flexibility	2.56	2.56	3.02	6.59
	Agmatine	0.00	-11.3	-12.4	-11.1
Fit to data	Rms residual	8.76	4.20	3.64	2.65
	Mean residual	5.57	2.04	0.14	0.0
	Correlation ( $r_P$ )	-0.403	0.613	0.617	0.666
	$R^2$	—	0.162	0.369	0.816
Cross-validated fit to data	$sPRESS$	—	5.09	4.55	3.70
	Mean residual	—	2.12	0.15	0.03
	$R_{CV}^2$	—	0.054	0.242	0.498
Fit to 67 protein- ligand complexes	Rms residual	7.93	8.63	10.0	27.87
	Mean residual	-0.91	-2.69	-6.27	-5.82

the comparison of the binding of ionic and non-ionic S1 moieties to trypsin and thrombin, where it is clear that the un-ionised molecules do not bind properly to trypsin and therefore may not exhibit the expected binding mode. This is a complex effect which would be difficult to predict *a priori* and might be primarily driven by the different local dielectric constant in the S1 pocket of trypsin and thrombin respectively. Such effects are probably too difficult to take into account in a normal drug design project where candidate designs of reasonable diversity must be quickly assessed with as much reliability as possible. It was these observations which drove us to explore the role of follow-up QSAR in conjunction with the application of empirical scoring functions on our structure based drug discovery projects.

Bayesian regression is particularly promising in drug discovery applications. It is normal to have a set of prior data about molecules binding to the receptor/enzyme of interest or related receptors and data is bound to accumulate during the progress of the project (if it experiences any success at all!). Commonly, QSAR can only be applied to the current data set and it is difficult to take into account prior information. In the above application we have shown how Bayesian regression can be used to fit new data sets while constraining the performance of the model against

prior information. In this case, it was believed that the original training set of crystallographically determined geometries and the associated ligand-receptor geometries provide useful and relevant prior knowledge about the way ligands bind to receptors. Beliefs about prior information, how useful and relevant it is and how trustworthy the prior data is relative to new assay data, are subjective. This subjectivity is often unappealing to those used to dealing with classical statistics. However we believe that ignoring the prior information or trying to treat all information on an equal footing will result in a model that is not as reliable. Subjective judgements are not eliminated by classical statistics since derived models are always judged with respect to prior information and prior beliefs about drug binding. We believe that it is often appropriate to put prior beliefs into the model and to deal rigorously with them using the Bayesian approach. The performance of the model can then be used to assess whether the prior beliefs have merit and add to the utility of the model.

The above application demonstrates the utility of this approach. For example, the application to trypsin produces a model with much better correlations than the original prediction (Figure 3 versus Figure 7). Classically, an alternative would be to run a regression directly on the new test sets using the descrip-

tors in Equation (1) augmented with a descriptor for the ionic interactions with Asp-189. However this leads to significant over-fitting. Another alternative which prevents overfitting would be to include the original data set in the new regression but such an approach is cumbersome and does not allow one to regard the data on the new designs as more important. Both the above alternatives are specific examples of Bayesian regression with different weights given to the prior information. Other classical alternatives exist but potentially suffer from the drawbacks of producing models with substantial over-fitting. It seems clear that the Bayesian approach, which allows one to apply a regression to new data whilst constraining it to fit previous information, will be preferable in many cases.

The application to thrombin also indicates the utility of the Bayesian method although in this case the omission of outliers and the specificity of the indicator variable make the resulting model less useful. The problem here is primarily a problem associated with the thrombin test set. There is simply not enough information to come to useful conclusions beyond the facts that agmatine is an important substituent and that some substituents do not work. Regressions which use more complicated variables arrive at complex ways of representing this, and thus are likely to be more misleading than the regression given here. As we get more data from our drug discovery projects on serine proteases, it is our intention to continuously update

Bayesian regression will be useful in drug discovery whenever meaningful prior estimates of activity are available. In the example presented here, computational methods offer good (albeit severely flawed) estimates of the binding affinity through the use of information on the target protein structure. Many other computational methods of estimating binding affinity could also be used [15, 16]. The Bayesian method could be used to bring in information from previous SAR on the enzyme or receptor of interest, and allow a project team to continuously refine the scoring function during structure based drug design applications. The method could also be useful in other areas of drug discovery. Information on the binding affinity against a set of enzymes can be used as prior information in applications against related enzymes. In a similar way, affinity fingerprints could be used as prior information in appropriate Bayesian QSAR applications [17]. Bayesian regression might also be useful in the calculation of physical properties on a particular chemical series of interest where there already exists good prior information on a large number of organic molecules [18].



## Conclusions

This paper has described the testing of a simple empirical scoring function on a set of molecular designs targetted at thrombin and produced by a *de novo* design program. The results demonstrate that the scoring function is useful but can not be relied on too heavily since the correlations obtained are poor. Changes in activity resulting from different analogues of a molecule are often too subtle to be predicted accurately. This highlights the need for some form of follow-up QSAR in drug design projects employing this type of empirical scoring function, perhaps through the addition of classical QSAR descriptors.

We have employed Bayesian regression in subsequent QSAR analysis of the molecular designs. To our knowledge, this is the first time Bayesian regression has been used in chemical QSAR analysis. Bayesian regression is a mathematically rigorous way of introducing prior information into new regressions. In our application, the prior information was derived from the original training set used to establish the empirical scoring function, i.e., a collection of crystal structures of ligands bound into their receptors and the associated binding affinities. The prior information is used to constrain the follow up regressions so that performance against the original training set is not significantly compromised. These results indicate that Bayesian regression is a useful and practical technique when relevant prior knowledge is available. It is anticipated that Bayesian regression will be helpful in other areas of chemical QSAR.

## Acknowledgements

The authors would like to thank David Frenkel and Bohdan Waskowycz for useful comments on an earlier version of this work.

## References

1. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., *J. Comput.-Aided Mol. Design*, 11 (1997) 425.
2. Böhm, H.-J., *J. Comput.-Aided Mol. Design*, 8 (1994) 243.
3. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., *J. Am. Chem. Soc.*, 118 (1996) 3959.
4. Jain, A.J., *J. Comput.-Aided Mol. Design*, 10 (1996) 427.
5. Murray, C.W., Clark, D.E., Auton, T.R., Firth, M.A., Li, J., Sykes, R.A., Waszkowycz, B., Westhead, D.R. and Young, S.C., *J. Comput.-Aided Mol. Design*, 11 (1997) 193.
6. Young, S.C., Auton, T.R., Clark, D.E., Li, J., Liebeschuetz, J.W., Lowe, R., Mahler, J., Martin, H., Morgan, P.J., Murray, C.W., Rimmer, A.D., Waszkowycz, B.W. and Westhead, D.R., *J. Med. Chem.*, submitted.
7. O'Hagan, A., *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, Wiley & Sons Inc., New York, NY, 1994.
8. Banner, D.W. and Hadvary, P., *J. Biol. Chem.*, 266 (1991) 20085.
9. Bode, W., Turk, D. and Stürzebecher, J., *Eur. J. Biochem.*, 193 (1990) 175.
10. DISCOVER, v 2.9.5, Molecular Simulations Inc., San Diego, CA, USA.
11. CFF95 Forcefield, implemented in DISCOVER 2.9.5., Molecular Simulations Inc., San Diego, CA, USA.
12. The NAG Fortran Library Manual, Mark 17, The Numerical Algorithms Group Ltd, Oxford, U.K., 1995.
13. Fersht, A.R., Shi, J., Knill-Jones, J., Lowe, D.M., Wilkinson, A.J., Blow, D.M., Brick, P., Carter, P., Waye, M.M.Y. and Winter, G., *Nature*, 314 (1985) 235.
14. Feng, D., Gardell, S.J., Dale Lewis, S., Bock, M.G., Chen, Z., Freidinger, R.M., Nayler-Olsen, A.M., Ramjit, H.M., Woltmann, R., Baskin, E.P., Lynch, J.J., Lucas, R., Shafer, J.A., Danacheck, K.B., Chen, I., Mao, S., Krueger, J.A., Hare, T.R., Mulichak, A.M. and Vacca, J.P., *J. Med. Chem.*, 40 (1997) 3726.
15. Ajay and Murcko, M.A., *J. Med. Chem.*, 38 (1995) 4953.
16. Aqvist, J., Medina, C. and Samuelsson, J., *Protein Eng.*, 3 (1994) 385.
17. Kauvar, L.M., Higgins, D.L., Villar, H.O., Sportsman, J.R., Engqvist-Goldstein, A.E., Bukar, R., Bauer, K.E., Dilley, H.M. and Rocke, D.M., *Chem. Biol.*, 2 (1995) 107.
18. Leo, A.J., *Chem. Rev.*, 93 (1993) 1281.

